Integrated Series in Information Systems

Volume 36

More information about this series at http://www.springer.com/series/6157

Shan Suthaharan

Machine Learning Models and Algorithms for Big Data Classification

Thinking with Examples for Effective Learning

Springer

Shan Suthaharan
Department of Computer Science
UNC Greensboro
Greensboro, NC, USA

ISSN 1571-0270 ISSN 2197-7968 (electronic)
Integrated Series in Information Systems
ISBN 978-1-4899-7640-6 ISBN 978-1-4899-7641-3 (eBook)
DOI 10.1007/978-1-4899-7641-3

Library of Congress Control Number: 2015950063

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

It is the quality of our work which will please God and not the quantity – Mahatma Gandhi

If you can't explain it simply, you don't understand it well enough – Albert Einstein

Preface

The interest in writing this book began at the IEEE International Conference on Intelligence and Security Informatics held in Washington, DC (June 11–14, 2012), where Mr. Matthew Amboy, the editor of *Business and Economics: OR and MS*, published by Springer Science+Business Media, expressed the need for a book on this topic, mainly focusing on a topic in data science field. The interest went even deeper when I attended the workshop conducted by Professor Bin Yu (Department of Statistics, University of California, Berkeley) and Professor David Madigan (Department of Statistics, Columbia University) at the Institute for Mathematics and its Applications, University of Minnesota on June 16–29, 2013.

Data science is one of the emerging fields in the twenty-first century. This field has been created to address the big data problems encountered in the day-to-day operations of many industries, including financial sectors, academic institutions, information technology divisions, health care companies, and government organizations. One of the important big data problems that needs immediate attention is in big data classifications. The network intrusion detection, public space intruder detection, fraud detection, spam filtering, and forensic linguistics are some of the practical examples of big data classification problems that require immediate attention.

We need significant collaboration between the experts in many disciplines, including mathematics, statistics, computer science, engineering, biology, and chemistry to find solutions to this challenging problem. Educational resources, like books and software, are also needed to train students to be the next generation of research leaders in this emerging research field. One of the current fields that brings the interdisciplinary experts, educational resources, and modern technologies under one roof is machine learning, which is a subfield of artificial intelligence.

Many models and algorithms for standard classification problems are available in the machine learning literature. However, a few of them are suitable for big data classification. Big data classification is dependent not only on the mathematical and software techniques but also on the computer technologies that help store, retrieve, and process the data with efficient scalability, accessibility, and computability features. One such recent technology is the distributed file system. A particular system

that has become popular and provides these features is the Hadoop distributed file system, which uses the modern techniques called MapReduce programming model (or a framework) with Mapper and Reducer functions that adopt the concept called the (key, value) pairs. The machine learning techniques such as the decision tree (a hierarchical approach), random forest (an ensemble hierarchical approach), and deep learning (a layered approach) are highly suitable for the system that addresses big data classification problems. Therefore, the goal of this book is to present some of the machine learning models and algorithms, and discuss them with examples.

The general objective of this book is to help readers, especially students and newcomers to the field of big data and machine learning, to gain a quick understanding of the techniques and technologies; therefore, the theory, examples, and programs (Matlab and R) presented in this book have been simplified, hardcoded, repeated, or spaced for improvements. They provide vehicles to test and understand the complicated concepts of various topics in the field. It is expected that the readers adopt these programs to experiment with the examples, and then modify or write their own programs toward advancing their knowledge for solving more complex and challenging problems.

The presentation format of this book focuses on simplicity, readability, and dependability so that both undergraduate and graduate students as well as new researchers, developers, and practitioners in this field can easily trust and grasp the concepts, and learn them effectively. The goal of the writing style is to reduce the mathematical complexity and help the vast majority of readers to understand the topics and get interested in the field. This book consists of four parts, with a total of 14 chapters. Part I mainly focuses on the topics that are needed to help analyze and understand big data. Part II covers the topics that can explain the systems required for processing big data. Part III presents the topics required to understand and select machine learning techniques to classify big data. Finally, Part IV concentrates on the topics that explain the scaling-up machine learning, an important solution for modern big data problems.

Greensboro, NC, USA Shan Suthaharan

Acknowledgements

The journey of writing this book would not have been possible without the support of many people, including my collaborators, colleagues, students, and family. I would like to thank all of them for their support and contributions toward the successful development of this book. First, I would like to thank Mr. Matthew Amboy (Editor, *Business and Economics: OR and MS*, Springer Science+Business Media) for giving me an opportunity to write this book. I would also like to thank both Ms. Christine Crigler (Assistant Editor) and Mr. Amboy for helping me throughout the publication process.

I am grateful to Professors Ratnasingham Shivaji (Head of the Department of Mathematics and Statistics at the University of North Carolina at Greensboro) and Fadil Santosa (Director of the Institute for Mathematics and its Applications at University of Minnesota) for the opportunities that they gave me to attend a machine learning workshop at the institute. Professors Bin Yu (Department of Statistics, University of California, Berkeley) and David Madigan (Department of Statistics, Columbia University) delivered an excellent short course on applied statistics and machine learning at the institute, and the topics covered in this course motivated me and equipped me with techniques and tools to write various topics in this book. My sincere thanks go to them. I would also like to thank Jinzhu Jia, Adams Bloniaz, and Antony Joseph, the members of Professor Bin Yu's research group at the Department of Statistics, University of California, Berkeley, for their valuable discussions in many machine learning topics.

My appreciation goes out to University of California, Berkeley, and University of North Carolina at Greensboro for their financial support and the research assignment award in 2013 to attend University of California, Berkeley as a Visiting scholar—this visit helped me better understand the deep learning techniques. I would also like to show my appreciation to Mr. Brent Ladd (Director of Education, Center for the Science of Information, Purdue University) and Mr. Robert Brown (Managing Director, Center for the Science of Information, Purdue University) for their support to develop a course on big data analytics and machine learning at University of North Carolina at Greensboro through a sub-award approved by the National Science Foundation. I am also thankful to Professor Richard Smith, Director of the

Statistical and Applied Mathematical Sciences Institute at North Carolina, for the opportunity to attend the workshops on low-dimensional structure in high-dimensional systems and to conduct research at the institute as a visiting research fellow during spring 2014. I greatly appreciate the resources that he provided during this visiting appointment. I also greatly appreciate the support and resources that the University of North Carolina at Greensboro provided during the development of this book.

The research work conducted with Professor Vaithilingam Jeyakumar and Dr. Guoyin Li at the University of New South Wales (Australia) helped me simplify the explanation of support vector machines. The technical report written by Michelle Dunbar under Professor Jeyakumar's supervision also contributed to the enhancement of the chapter on support vector machines. I would also like to express my gratitude to Professors Sat Gupta, Scott Richter, and Edward Hellen for sharing their knowledge of some of the statistical and mathematical techniques. Professor Steve Tate's support and encouragement, as the department head and as a colleague, helped me engage in this challenging book project for the last three semesters. My sincere gratitude also goes out to Professor Jing Deng for his support and engagement in some of my research activities.

My sincere thanks also go to the following students who recently contributed directly or indirectly to my research and knowledge that helped me develop some of the topics presented in this book: Piyush Agarwal, Mokhaled Abd Allah, Michelle Bayait, Swarna Bonam, Chris Cain, Tejo Sindhu Chennupati, Andrei Craddock, Luning Deng, Anudeep Katangoori, Sweta Keshpagu, Kiranmayi Kotipalli, Varnika Mittal, Chitra Reddy Musku, Meghana Narasimhan, Archana Polisetti, Chadwik Rabe, Naga Padmaja Tirumal Reddy, Tyler Wendell, and Sumanth Reddy Yanala.

Finally, I would like to thank my wife, Manimehala Suthaharan, and my lovely children, Lovepriya Suthaharan, Praveen Suthaharan, and Prattheeba Suthaharan, for their understanding, encouragement, and support which helped me accomplish this project. This project would not have been completed successfully without their support.

Greensboro, NC, USA Shan Suthaharan
June 2015

About the Author

Shan Suthaharan is a Professor of Computer Science at the University of North Carolina at Greensboro (UNCG), North Carolina, USA. He also serves as the Director of Undergraduate Studies at the Department of Computer Science at UNCG. He has more than 25 years of university teaching and administrative experience and has taught both undergraduate and graduate courses. His aspiration is to educate and train students so that they can prosper in the computer field by understanding current real-world and complex problems, and develop efficient techniques and technologies. His current teaching interests include big data analytics and machine learning, cryptography and network security, and computer networking and analysis. He earned his doctorate in Computer Science from Monash University, Australia. Since then, he has been actively working on disseminating his knowledge and experience through teaching, advising, seminars, research, and publications.

Dr. Suthaharan enjoys investigating real-world, complex problems, and developing and implementing algorithms to solve those problems using modern technologies. The main theme of his current research is the signature discovery and event detection for a secure and reliable environment. The ultimate goal of his research is to build a secure and reliable environment using modern and emerging technologies. His current research primarily focuses on the characterization and detection of environmental events, the exploration of machine learning techniques, and the development of advanced statistical and computational techniques to discover key signatures and detect emerging events from structured and unstructured big data.

Dr. Suthaharan has authored or co-authored more than 75 research papers in the areas of computer science, and published them in international journals and refereed conference proceedings. He also invented a key management and encryption technology, which has been patented in Australia, Japan, and Singapore. He also received visiting scholar awards from and served as a visiting researcher at the University of Sydney, Australia; the University of Melbourne, Australia; and the University of California, Berkeley, USA. He was a senior member of the Institute of Electrical and Electronics Engineers, and volunteered as an elected chair of the Central North Carolina Section twice. He is a member of Sigma Xi, the Scientific Research Society and a Fellow of the Institution of Engineering and Technology.

Contents

1 Science of Information 1
1.1 Data Science 1
1.1.1 Technological Dilemma 2
1.1.2 Technological Advancement 2
1.2 Big Data Paradigm 3
1.2.1 Facts and Statistics of a System 3
1.2.2 Big Data Versus Regular Data 5
1.3 Machine Learning Paradigm 7
1.3.1 Modeling and Algorithms 7
1.3.2 Supervised and Unsupervised 7
1.4 Collaborative Activities 10
1.5 A Snapshot 10
1.5.1 The Purpose and Interests 10
1.5.2 The Goal and Objectives 11
1.5.3 The Problems and Challenges 11
Problems 11
References 12

Part I Understanding Big Data

2 Big Data Essentials 17
2.1 Big Data Analytics 17
2.1.1 Big Data Controllers 18
2.1.2 Big Data Problems 19
2.1.3 Big Data Challenges 19
2.1.4 Big Data Solutions 20
2.2 Big Data Classification 20
2.2.1 Representation Learning 21
2.2.2 Distributed File Systems 22
2.2.3 Classification Modeling 23
2.2.4 Classification Algorithms 25

2.3 Big Data Scalability 26
2.3.1 High-Dimensional Systems 27
2.3.2 Low-Dimensional Structures 27
Problems 28
References 28

3 Big Data Analytics 31
3.1 Analytics Fundamentals 31
3.1.1 Research Questions 32
3.1.2 Choices of Data Sets 33
3.2 Pattern Detectors 34
3.2.1 Statistical Measures 34
3.2.2 Graphical Measures 38
3.2.3 Coding Example 41
3.3 Patterns of Big Data 44
3.3.1 Standardization: A Coding Example 47
3.3.2 Evolution of Patterns 49
3.3.3 Data Expansion Modeling 51
3.3.4 Deformation of Patterns 62
3.3.5 Classification Errors 66
3.4 Low-Dimensional Structures 67
3.4.1 A Toy Example 67
3.4.2 A Real Example 69
Problems 73
References 74

Part II Understanding Big Data Systems

4 Distributed File System 79
4.1 Hadoop Framework 79
4.1.1 Hadoop Distributed File System 80
4.1.2 MapReduce Programming Model 81
4.2 Hadoop System 81
4.2.1 Operating System 82
4.2.2 Distributed System 82
4.2.3 Programming Platform 83
4.3 Hadoop Environment 83
4.3.1 Essential Tools 84
4.3.2 Installation Guidance 85
4.3.3 RStudio Server 93
4.4 Testing the Hadoop Environment 94
4.4.1 Standard Example 94
4.4.2 Alternative Example 95

4.5 Multinode Hadoop 95
4.5.1 Virtual Network 96
4.5.2 Hadoop Setup 96
Problems 97
References 97

5 MapReduce Programming Platform 99
5.1 MapReduce Framework 99
5.1.1 Parametrization 100
5.1.2 Parallelization 101
5.2 MapReduce Essentials 102
5.2.1 Mapper Function 102
5.2.2 Reducer Function 103
5.2.3 MapReduce Function 104
5.2.4 A Coding Example 104
5.3 MapReduce Programming 107
5.3.1 Naming Convention 107
5.3.2 Coding Principles 108
5.3.3 Application of Coding Principles 110
5.4 File Handling in MapReduce 113
5.4.1 Pythagorean Numbers 114
5.4.2 File Split Example 115
5.4.3 File Split Improved 116
Problems 118
References 118

Part III Understanding Machine Learning

6 Modeling and Algorithms 123
6.1 Machine Learning 123
6.1.1 A Simple Example 124
6.1.2 Domain Division Perspective 125
6.1.3 Data Domain 128
6.1.4 Domain Division 129
6.2 Learning Models 130
6.2.1 Mathematical Models 132
6.2.2 Hierarchical Models 134
6.2.3 Layered Models 135
6.2.4 Comparison of the Models 135
6.3 Learning Algorithms 140
6.3.1 Supervised Learning 140
6.3.2 Types of Learning 141
Problems 142
References 142

7 Supervised Learning Models 145
7.1 Supervised Learning Objectives 145
7.1.1 Parametrization Objectives 146
7.1.2 Optimization Objectives 148
7.2 Regression Models 150
7.2.1 Continuous Response 151
7.2.2 Theory of Regression Models 151
7.3 Classification Models 160
7.3.1 Discrete Response 160
7.3.2 Mathematical Models 162
7.4 Hierarchical Models 166
7.4.1 Decision Tree 167
7.4.2 Random Forest 167
7.5 Layered Models 170
7.5.1 Shallow Learning 171
7.5.2 Deep Learning 177
Problems 179
References 180

8 Supervised Learning Algorithms 183
8.1 Supervised Learning 183
8.1.1 Learning 185
8.1.2 Training 186
8.1.3 Testing 188
8.1.4 Validation 190
8.2 Cross-Validation 192
8.2.1 Tenfold Cross-Validation 193
8.2.2 Leave-One-Out 193
8.2.3 Leave-p-Out 194
8.2.4 Random Subsampling 195
8.2.5 Dividing Data Sets 195
8.3 Measures 196
8.3.1 Quantitative Measure 197
8.3.2 Qualitative Measure 198
8.4 A Simple 2D Example 202
Problems 204
References 205

9 Support Vector Machine 207
9.1 Linear Support Vector Machine 207
9.1.1 Linear Classifier: Separable Linearly 208
9.1.2 Linear Classifier: Nonseparable Linearly 218
9.2 Lagrangian Support Vector Machine 219
9.2.1 Modeling of LSVM 219
9.2.2 Conceptualized Example 219
9.2.3 Algorithm and Coding of LSVM 220

9.3 Nonlinear Support Vector Machine . . . 223
9.3.1 Feature Space . . . 224
9.3.2 Kernel Trick . . . 224
9.3.3 SVM Algorithms on Hadoop . . . 227
9.3.4 Real Application . . . 233
Problems . . . 234
References . . . 235

10 Decision Tree Learning . . . 237
10.1 The Decision Tree . . . 237
10.1.1 A Coding Example—Classification Tree . . . 241
10.1.2 A Coding Example—Regression Tree . . . 244
10.2 Types of Decision Trees . . . 245
10.2.1 Classification Tree . . . 246
10.2.2 Regression Tree . . . 247
10.3 Decision Tree Learning Model . . . 248
10.3.1 Parametrization . . . 248
10.3.2 Optimization . . . 249
10.4 Quantitative Measures . . . 250
10.4.1 Entropy and Cross-Entropy . . . 250
10.4.2 Gini Impurity . . . 252
10.4.3 Information Gain . . . 255
10.5 Decision Tree Learning Algorithm . . . 256
10.5.1 Training Algorithm . . . 257
10.5.2 Validation Algorithm . . . 263
10.5.3 Testing Algorithm . . . 263
10.6 Decision Tree and Big Data . . . 266
10.6.1 Toy Example . . . 266
Problems . . . 268
References . . . 269

Part IV Understanding Scaling-Up Machine Learning

11 Random Forest Learning . . . 273
11.1 The Random Forest . . . 273
11.1.1 Parallel Structure . . . 274
11.1.2 Model Parameters . . . 275
11.1.3 Gain/Loss Function . . . 276
11.1.4 Bootstrapping and Bagging . . . 276
11.2 Random Forest Learning Model . . . 278
11.2.1 Parametrization . . . 279
11.2.2 Optimization . . . 279
11.3 Random Forest Learning Algorithm . . . 279
11.3.1 Training Algorithm . . . 280
11.3.2 Testing Algorithm . . . 283

11.4 Random Forest and Big Data 284
11.4.1 Random Forest Scalability 284
11.4.2 Big Data Classification 284
Problems 287
References 288

12 Deep Learning Models 289
12.1 Introduction 289
12.2 Deep Learning Techniques 291
12.2.1 No-Drop Deep Learning 291
12.2.2 Dropout Deep Learning 291
12.2.3 Dropconnect Deep Learning 292
12.2.4 Gradient Descent 293
12.2.5 A Simple Example 297
12.2.6 MapReduce Implementation 298
12.3 Proposed Framework 301
12.3.1 Motivation 301
12.3.2 Parameters Mapper 301
12.4 Implementation of Deep Learning 303
12.4.1 Analysis of Domain Divisions 303
12.4.2 Analysis of Classification Accuracies 303
12.5 Ensemble Approach 305
Problems 306
References 306

13 Chandelier Decision Tree 309
13.1 Unit Circle Algorithm 309
13.1.1 UCA Classification 310
13.1.2 Improved UCA Classification 311
13.1.3 A Coding Example 312
13.1.4 Drawbacks of UCA 315
13.2 Unit Circle Machine 315
13.2.1 UCM Classification 315
13.2.2 A Coding Example 316
13.2.3 Drawbacks of UCM 318
13.3 Unit Ring Algorithm 318
13.3.1 A Coding Example 319
13.3.2 Unit Ring Machine 321
13.3.3 A Coding Example 321
13.3.4 Drawbacks of URM 323
13.4 Chandelier Decision Tree 323
13.4.1 CDT-Based Classification 324
13.4.2 Extension to Random Chandelier 328
Problems 328
References 328

14 Dimensionality Reduction 329
14.1 Introduction 329
14.2 Feature Hashing Techniques 330
14.2.1 Standard Feature Hashing 331
14.2.2 Flagged Feature Hashing 331
14.3 Proposed Feature Hashing 332
14.3.1 Binning and Mitigation 332
14.3.2 Mitigation Justification 333
14.3.3 Toy Example 333
14.4 Simulation and Results 334
14.4.1 A Matlab Implementation 334
14.4.2 A MapReduce Implementation 337
14.5 Principal Component Analysis 340
14.5.1 Eigenvector 341
14.5.2 Principal Components 343
14.5.3 The Principal Directions 346
14.5.4 A 2D Implementation 348
14.5.5 A 3D Implementation 350
14.5.6 A Generalized Implementation 352
Problems 354
References 354

Index 357